(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102737027 A
(43)申请公布日 2012.10.17

(21)申请号 201110082369.8

(22)申请日 2011.04.01

(71)申请人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室

(72)发明人 王亮 文勖 焦峰 王锐 付剑波 许春林 石一峰 刘晓云

(74)专利代理机构 广州华进联合专利商标代理有限公司 44224
代理人 何平 曾旻辉

(51)Int.Cl.
G06F 17/30(2006.01)

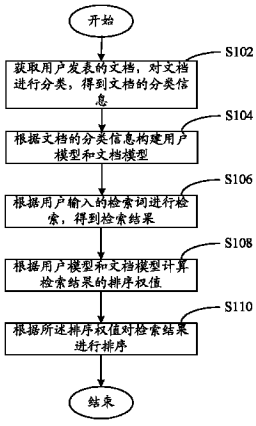
权利要求书 2 页 说明书 9 页 附图 3 页

(54)发明名称

个性化搜索方法及系统

(57)摘要

一种个性化搜索方法，包括以下步骤：获取用户发表的文档，对文档进行分类，得到文档的分类信息；根据文档的分类信息构建用户模型和文档模型；根据用户输入的检索词进行检索，得到检索结果；根据所述用户模型和文档模型计算所述检索结果排序权值；根据所述排序权值对所述检索结果进行排序。采用上述方法，构建的用户模型和文档模型的维度较低，实现起来简单，从而能够提高运行性能。此外，还提供了一种个性化搜索系统。



CN 102737027 A

1. 一种个性化搜索方法，包括以下步骤：

获取用户发表的文档，对文档进行分类，得到文档的分类信息；

根据文档的分类信息构建用户模型和文档模型；

根据用户输入的检索词进行检索，得到检索结果；

根据所述用户模型和文档模型计算所述检索结果的排序权值；

根据所述排序权值对所述检索结果进行排序。

2. 根据权利要求 1 所述的个性化搜索方法，其特征在于，所述根据文档的分类信息构建用户模型和文档模型的步骤为：

获取用户发表的文档的分类概率及检索词的分类概率；

构建个人模型，所述个人模型为由用户发表的文档的分类概率组成的向量，构建大众模型，所述大众模型为由检索词的分类概率组成的向量，将所述个人模型与大众模型进行线性叠加，得到用户模型；

构建文档模型，所述文档模型为由文档属于各分类的概率组成的向量。

3. 根据权利要求 2 所述的个性化搜索方法，其特征在于，采用如下公式构建所述用户模型：

$P(people\_social) = a \times P(query) + (1-a)P(people), 0 \leqslant a \leqslant 1$

其中，P(people_social) 为用户模型，P(query) 为大众模型，P(people) 为个人模型，a 为用户活跃度指数；

所述用户活跃度指数的计算公式为：

$$a = \begin{cases} \dfrac{N}{2 \times N1}, N < 2 \times N1 \\ 1, N \geq 2 \times N1 \end{cases}$$

其中，N 为一个用户发表的文档总数，N1 为所有用户平均发表的文档数。

4. 根据权利要求 1 所述的个性化搜索方法，其特征在于，所述根据用户模型和文档模型计算所述检索结果的排序权值的步骤为：

获取登录用户的用户模型和所述检索结果中每个文档的文档模型；

计算所述登录用户的用户模型和所述文档模型的第一相似度；

获取检索结果中每个文档的作者的用户模型，计算所述作者的用户模型与所述登录用户的用户模型的第二相似度；

将所述第一相似度与第二相似度进行线性叠加，得到所述排序权值。

5. 根据权利要求 1 所述的个性化搜索方法，其特征在于，所述方法还包括对用户模型进行更新的步骤，具体是：

设置线性衰减函数、统计周期和统计时间段；

在所述统计时间段内统计每个统计周期内用户发表的各类文档数；

根据所述线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数；

根据所述折算后的用户发表的各类文档数构建用户模型。

6. 一种个性化搜索系统，其特征在于，包括：

文档分类模块，用于获取用户发表的文档，并对文档进行分类，得到文档的分类信息；

用户模型构建模块，用于获取所述文档分类信息，根据所述文档分类信息构建用户模型；

文档模型构建模块，用于获取所述文档分类信息，根据所述文档分类信息构建文档模型；

检索模块，用于根据用户输入的检索词进行检索，得到检索结果；

排序权值计算模块，用于根据所述用户模型和文档模型计算所述检索结果的排序权值；

排序模块，用于根据所述排序权值对所述检索结果进行排序。

7. 根据权利要求 6 所述的个性化搜索系统，其特征在于，所述用户模型构建模块用于获取用户发表的文档的分类概率及检索词的分类概率，构建个人模型，所述个人模型为由用户发表的文档的分类概率组成的向量，构建大众模型，所述大众模型为由检索词的分类概率组成的向量，将所述个人模型与大众模型进行线性叠加，得到用户模型；

所述文档模型构建模块用于构建所述文档模型，所述文档模型为由文档属于各分类的概率组成的向量。

8. 根据权利要求 7 所述的个性化搜索系统，其特征在于，所述用户模型构建模块按如下公式构建所述用户模型：

$P(people\_social) = a \times P(query) + (1-a)P(people), 0 \leqslant a \leqslant 1$

其中，P(people_social) 为用户模型，P(query) 为大众模型，P(people) 为个人模型，a 为用户活跃度指数；

所述系统还包括用户活跃度指数计算模块，所述用户活跃度指数计算模块计算所述用户活跃度指数的计算公式为：

$$a=\begin{cases}\dfrac{N}{2\times N1}, N<2\times N1\\ 1, N\geq 2\times N1\end{cases}$$

其中，N 为一个用户发表的文档总数，N1 为所有用户平均发表的文档数。

9. 根据权利要求 6 所述的个性化搜索系统，其特征在于，所述排序权值计算模块包括：

查询单元，用于获取登录用户的用户模型；

相似度计算单元，用于获取检索结果中每个文档的文档模型及每个文档的作者的用户模型，计算所述登录用户的用户模型和所述文档模型的第一相似度，以及计算所述作者的用户模型与所述登录用户的用户模型的第二相似度；

线性叠加单元，用于将所述第一相似度和第二相似度进行线性叠加，得到排序权值。

10. 根据权利要求 6 所述的个性化搜索系统，其特征在于，所述用户模型构建模块包括更新模块，所述更新模块具体包括：

设置单元，用于设置线性衰减函数、统计周期和统计时间段；

统计单元，用于在所述统计时间段内统计每个统计周期内用户发表的各类文档数；

折算单元，用于根据所述线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数；

用户模型构建单元，用于根据所述折算后的用户发表的各类文档数构建用户模型。

个性化搜索方法及系统

【技术领域】

[0001] 本发明涉及搜索技术，尤其涉及一种个性化搜索方法及系统。

【背景技术】

[0002] 个性化搜索是一种信息搜索方式，相对于普通搜索方式其考虑了用户的区别，利用用户信息对搜索结果进行修改或过滤，以得到更符合用户个性化需求的搜索结果。个性化搜索的基本方法是将用户输入的关键词和用户的个人偏好联系起来进行查询，从而得到用户最可能需要的信息显示在最前面。

[0003] 个性化搜索需要解决两个问题，一是如何构建用户模型，二是如何对搜索结果进行重新排序。构建用户模型需要先采集用户信息，包括用户注册时提供的职业、毕业院校、兴趣爱好等个人信息及用户访问日志等。传统的个性化搜索方式中，是基于 IP、Cookie 等方式来采集用户信息。根据采集到的用户信息构建用户模型，传统的个性化搜索方式中，通常采用基于内容的方法来构建用户模型，利用用户信息中的特征词来表示用户的兴趣，构建的是基于特征词的向量空间模型。然而，这样所构建的用户模型维度过高，一般都是几万维，而个性化搜索都需要在线实时计算，在用户模型维度过高的情况下，实现起来非常困难。

【发明内容】

[0004] 基于此，有必要提供一种实现简单、能提高运行性能的个性化搜索方法。

[0005] 一种个性化搜索方法，包括以下步骤：获取用户发表的文档，对文档进行分类，得到文档的分类信息；根据文档的分类信息构建用户模型和文档模型；根据用户输入的检索词进行检索，得到检索结果；根据所述用户模型和文档模型计算所述检索结果的排序权值；根据所述排序权值对所述检索结果进行排序。

[0006] 优选的，所述根据文档的分类信息构建用户模型和文档模型的步骤为：获取用户发表的文档的分类概率及检索词的分类概率；构建个人模型，所述个人模型为由用户发表的文档的分类概率组成的向量，构建大众模型，所述大众模型为由检索词的分类概率组成的向量，将所述个人模型与大众模型进行线性叠加，得到用户模型；构建文档模型，所述文档模型为由文档属于各分类的概率组成的向量。

[0007] 优选的，采用如下公式构建所述用户模型：

[0008] $P(people\_social) = a \times P(query) + (1-a)P(people)$，$0 \leqslant a \leqslant 1$

[0009] 其中，$P(people\_social)$ 为用户模型，$P(query)$ 为大众模型，$P(people)$ 为个人模型，a 为用户活跃度指数；

[0010] 所述用户活跃度指数的计算公式为：

[0011] $$a=\begin{cases}\dfrac{N}{2\times N1}, N<2\times N1\\ 1, N\geq 2\times N1\end{cases}$$

[0012] 其中，N 为一个用户发表的文档总数，N1 为所有用户平均发表的文档数。

[0013] 优选的，所述根据用户模型和文档模型计算排序权值的步骤为：获取登录用户的用户模型和检索结果中每个文档的文档模型；计算所述登录用户的用户模型和所述文档模型的第一相似度；获取检索结果中每个文档的作者的用户模型，计算所述作者的用户模型与所述登录用户的用户模型的第二相似度；将所述第一相似度与第二相似度进行线性叠加，得到所述排序权值。

[0014] 优选的，所述方法还包括对用户模型进行更新的步骤，具体是：设置线性衰减函数、统计周期和统计时间段；在所述统计时间段内统计每个统计周期内用户发表的各类文档数；根据所述线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数；根据所述折算后的用户发表的各类文档数构建用户模型。

[0015] 此外，还有必要提供一种实现简单、能提高运行性能的个性化搜索系统。

[0016] 一种个性化搜索系统，包括：文档分类模块，用于获取用户发表的文档，并对文档进行分类，得到文档的分类信息；用户模型构建模块，用于获取所述文档分类信息，根据所述文档分类信息构建用户模型；文档模型构建模块，用于获取所述文档分类信息，根据所述文档分类信息构建文档模型；检索模块，用于根据用户输入的检索词进行检索，得到检索结果；排序权值计算模块，用于根据所述用户模型和文档模型计算所述检索结果的排序权值；排序模块，用于根据所述排序权值对所述检索结果进行排序。

[0017] 优选的，所述用户模型构建模块用于获取用户发表的文档的分类概率及检索词的分类概率，构建个人模型，所述个人模型为由用户发表的文档的分类概率组成的向量，构建大众模型，所述大众模型为由检索词的分类概率组成的向量，将所述个人模型与大众模型进行线性叠加，得到用户模型；所述文档模型构建模块用于构建所述文档模型，所述文档模型为由文档属于各分类的概率组成的向量。

[0018] 优选的，所述用户模型构建模块按如下公式构建所述用户模型：

[0019] $P(people\_social) = a\times P(query)+(1-a)P(people), 0\leq a\leq 1$

[0020] 其中，P(people_social) 为用户模型，P(query) 为大众模型，P(people) 为个人模型，a 为用户活跃度指数；

[0021] 所述系统还包括用户活跃度指数计算模块，所述用户活跃度指数计算模块计算所述用户活跃度指数的计算公式为：

[0022] $$a=\begin{cases}\dfrac{N}{2\times N1}, N<2\times N1\\ 1, N\geq 2\times N1\end{cases}$$

[0023] 其中，N 为一个用户发表的文档总数，N1 为所有用户平均发表的文档数。

[0024] 优选的，所述排序权值计算模块包括：查询单元，用于获取登录用户的用户模型；相似度计算单元，用于获取检索结果中每个文档的文档模型及每个文档的作者的用户模

型，计算所述登录用户的用户模型和所述文档模型的第一相似度，以及计算所述作者的用户模型与所述登录用户的用户模型的第二相似度；线性叠加单元，用于将所述第一相似度和第二相似度进行线性叠加，得到排序权值。

[0025] 优选的，所述用户模型构建模块包括更新模块，所述更新模块具体包括：设置单元，用于设置线性衰减函数、统计周期和统计时间段；统计单元，用于在所述统计时间段内统计每个统计周期内用户发表的各类文档数；折算单元，用于根据所述线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数；用户模型构建单元，用于根据所述折算后的用户发表的各类文档数构建用户模型。

[0026] 上述个性化搜索方法和系统，根据文档分类构建用户模型和文档模型，再根据构建的用户模型和文档模型计算检索结果的排序权值，按照排序权值对检索结果进行重新排序。由于分类维度通常在十几维以内，因此构建的用户模型和文档模型的维度较低，实现起来简单，从而能够提高运行性能。

【附图说明】

[0027] 图 1 为一个实施例中个性化搜索方法的流程图；

[0028] 图 2 为图 1 中计算排序权值的方法流程图；

[0029] 图 3 为一个实施例中对用户模型进行更新的方法流程图；

[0030] 图 4 为一个实施例中个性化搜索系统的结构示意图；

[0031] 图 5 为图 4 中的排序权值计算模块的结构示意图；

[0032] 图 6 为一个实施例中更新模块的结构示意图。

【具体实施方式】

[0033] 如图 1 所示，一种个性化搜索方法，包括以下步骤：

[0034] 步骤 S102，获取用户发表的文档，对文档进行分类，得到文档的分类信息。

[0035] 用户发表的文档包括用户通过网络社区发表的各种日志、博客、评论等。用户通过网络社区发表的文档最能展示用户的兴趣，根据这些文档构建的用户模型更准确。可采用传统的文本自动分类方法对用户发表的文档进行分类。用户发表的文档及对这些文档的分类信息存储在搜索引擎数据库中。

[0036] 步骤 S104，根据文档的分类信息构建用户模型和文档模型。

[0037] 在一个实施例中，根据文档的分类信息构建的用户模型为个人模型和大众模型的线性叠加。其中，个人模型为由用户发表的文档的分类概率组成的向量。如有 n 个分类，那么个人模型就是一个 n 维的向量，其向量的各个元素为用户发表的文档的分类概率。例如，个人模型可以表示为：

[0038] $P(people) = (y_1, y_2, \cdots, y_n)$

[0039] 其中，y_i 为用户发表的一类文档的分类概率。如设置两个类别，体育类文档和数码科技类文档。用户共发布体育类文档 20 篇，发表数码科技类文档 80 篇。则用户的个人模型为 $P(people) = (0.2, 0.8)$。

[0040] 由于部分用户发表的文档可能很少，其个人模型的可信度并不高，并且在个性化搜索实施初期，很多用户需要一个初步接受的过程，因此需要设置大众模型。大众模型为由

检索词的分类概率组成的向量。由于检索词很短，可采用传统的贝叶斯自动分类方法估算其分类概率。例如，大众模型可以表示为：

[0041] $P(query) = (x_1, x_2, \cdots, x_n)$

[0042] 其中，x_i 为检索词属于文档各分类的概率。如设置了体育和数码科技两个类别，检索词“苹果”是体育类的概率为 0.2，是数码科技类的概率为 0.8，则用户的大众模型为 $P(query) = (0.2, 0.8)$。

[0043] 该实施例中，根据个人模型和大众模型构建的用户模型可表示为：

[0044] $P(people\_social) = a \times P(query) + (1-a)P(people)$，$0 \leqslant a \leqslant 1$

[0045] 其中，$P(people\_social)$ 为用户模型，$P(query)$ 为大众模型，$P(people)$ 为个人模型，a 为用户活跃度指数。

[0046] 用户活跃度指数用来判断个人模型的可信度。有的用户仅发表几篇文档，而有的用户发表上千篇文档，用户活跃度指数能够用来综合考虑大众兴趣和个人兴趣。用户发表的文档总数太少，则其兴趣主要以大众兴趣为主，用户发表的文档总数很多，则增加个人模型的权重，因此用户活跃度指数影响个人模型和大众模型的线性叠加参数。

[0047] 在一个优选的实施例中，所述用户活跃度指数的计算公式为：

[0048] $$a = \begin{cases} \dfrac{N}{2 \times N1}, N < 2 \times N1 \\ 1, N \geq 2 \times N1 \end{cases}$$

[0049] 其中，N 为一个用户发表的文档总数，N1 为所有用户平均发表的文档数。即：如果一个用户发表的文档数超过用户平均发表文档数的 2 倍，则用户活跃度指数为 1，认为其个人模型是可信的。如果小于用户平均发表文档数的 2 倍，则计算用户发表的文档数与 2 倍的用户平均发表文档数的比值，作为个人模型和大众模型进行线性叠加计算的参数，计算用户模型。

[0050] 在一个实施例中，构建文档模型，该文档模型为由文档属于各分类的概率组成的向量。如有 n 个分类，则文档模型就是一个 n 维的向量，其向量的各个元素为文档属于各个分类的概率。例如可以表示为：

[0051] $P(page) = (z_1, z_2, \cdots, z_n)$

[0052] 其中，z_i 为文档属于各分类的概率。对于长文本（如博客等），可采用传统的支持向量机中的逻辑回归方法估算文档的分类概率，对于短文本（如微博等），可采用传统的贝叶斯自动分类方法估算其分类概率。

[0053] 步骤 S106，根据用户输入的检索词进行检索，得到检索结果。

[0054] 步骤 S108，根据用户模型和文档模型计算检索结果的排序权值。

[0055] 步骤 S110，根据排序权值对检索结果进行排序。

[0056] 根据排序权值对检索结果进行排序后，排序靠前的检索结果更能反映用户的兴趣，所提供的个性化搜索结果更能满足用户的需求。由于根据文档分类来构建用户模型和文档模型，分类维度较低，使得构建的用户模型和文档模型的维度也较低，实现起来简单，能够提高运行性能。

[0057] 此外，基于分类方法来构建用户模型和文档模型，容易设置多级分类，易于扩展，

也不会因为分类的增加而降低运行性能。并且，根据分类概率构建的用户模型能更好的体现用户兴趣的多样性和变化，重新排序后的检索结果更能满足用户需求。用户模型和个人模型为分类概率所组成的向量，所构建的模型提供的仅仅是维度很低的数字，涉及宏观上的兴趣类别，不会涉及到用户的隐私信息，易于被用户所接受。

[0058] 在一个实施例中，如图 2 所示，步骤 S108 的具体过程为：

[0059] 步骤 S202，获取登录用户的用户模型和检索结果中每个文档的文档模型。

[0060] 根据登录用户的标识号可查找到登录用户的用户模型。根据用户输入的检索词进行检索，获取得到的检索结果中每个文档的文档模型。

[0061] 步骤 S204，计算登录用户的用户模型和文档模型的第一相似度。

[0062] 该实施例中，计算第一相似度即计算用户的用户模型与文档模型的余弦距离，计算公式为：rank1 = sim(P(people)，P(page)) = cos<P(people)，P(page)>。

[0063] 步骤 S206，获取检索结果中每个文档的作者的用户模型，计算作者的用户模型与登录用户的用户模型的第二相似度。

[0064] 根据检索结果中每个文档的作者的标识获取到对应的用户模型。计算第二相似度即计算每个文档的作者的用户模型与登录用户的用户模型之间的余弦距离，计算公式为：rank2 = sim(P(user)，P(author)) = cos<P(user)，P(author)>，其中，P(user) 为登录用户的用户模型，P(author) 为每个文档的作者的用户模型。

[0065] 在一个优选的实施例中，步骤 S206 还包括计算专家指数的步骤。专家指数用于第二相似度的计算中，其计算公式为：

[0066] $$\exp(M)=\frac{M}{M1}$$

[0067] 其中，M 为用户发表的某类文档的总数，M1 为发表该类别文档最多的用户发布的该类文档的数目。

[0068] 在一个实施例中，计算专家指数后，每个文档的作者的用户模型与登录用户的用户模型的第二相似度的计算公式则为：

[0069] rank2 = cos<P(user)，P(author)>*exp(M)

[0070] 也可以采用其他方式修正第二相似度的计算，例如，利用专家指数对第二相似度进行加权处理等。

[0071] 步骤 S208，将第一相似度与第二相似度进行线性叠加，得到排序权值。

[0072] 该实施例中，按照如下公式计算排序权值：

[0073] rank = b×rank1+(1-b)rank2，0 ≤ b ≤ 1

[0074] 其中，b 为经验参数。

[0075] 在另一个实施例中，上述个性化搜索方法还包括对用户模型进行更新的步骤。由于文档本身的分类变化较少，因此文本模型通常不需要更新。而用户的兴趣会随着时间变化，因此在一定时间周期内需要对用户模型进行更新。

[0076] 该实施例中，如图 3 所示，对用户模型进行更新的具体过程如下：

[0077] 步骤 S302 中，设置线性衰减函数、统计周期和统计时间段。

[0078] 由于用户模型是按照其发表文档的分类概率构建的，将各时段发表的文档的篇数按照距离当前的时间进行衰减，然后统一计算衰减后的各类别的文档数，根据衰减后的各

类别的文档数构建用户模型，则完成了用户模型的更新。

[0079] 步骤 S304 中，在统计时间段内统计每个统计周期内用户发表的各类文档数。

[0080] 步骤 S306 中，根据线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数。

[0081] 在一个优选的实施例中，设置衰减函数为 f(t) = t，表示用户兴趣随时间线性衰减。设置统计周期为月，12 个月为一个统计时间段。统计 12 个月每个月发表的各类文档数，按照距当前的时间进行折算，计算折算后的用户发表的各类文档数。

[0082] 该实施例中，采用如下公式进行折算：

[0083] $$Z = \sum_{t=1}^{T} (z_t / f(t))$$

[0084] 其中，Z 表示折算后作者发表的某个类别的文档数，t 表示文档发表日期距离当前的时间，z_t 表示 t 时段作者发表的某个类别的文档数，f(t) 为衰减函数。

[0085] 例如，一个用户在统计时间段内，第一个月发表体育类文档数 z1，第二个月发表体育类文档数 z2，...，当前的月份，也就是第 12 个月发表体育文档数为 z12，则该用户发布的体育类文档折算后的数目为 $Z = \frac{z1}{12} + \frac{z2}{11} + \frac{z3}{10} + \cdots + \frac{z11}{2} + z12$。

[0086] 步骤 S308 中，根据折算后的用户发表的各类文档数构建当前的用户模型。

[0087] 如图 4 所示，一种个性化搜索系统，包括文档分类模块 100、搜索引擎数据库 200、用户模型构建模块 300、文档模型构建模块 400、检索模块 500、排序权值计算模块 600 和排序模块 700，其中：

[0088] 文档分类模块 100 用于获取用户发表的文档，并对文档进行分类，得到文档的分类信息。

[0089] 用户发表的文档包括用户通过网络社区发表的各种日志、博客、评论等。用户通过网络社区发表的文档最能展示用户的兴趣，根据这些文档构建的用户模型更准确。可采用传统的文本自动分类方法对用户发表的文档进行分类。用户发表的文档及对这些文档的分类信息可存储在搜索引擎数据库 200 中。

[0090] 搜索引擎数据库 200 用于存储用户发表的文档和文档分类信息。

[0091] 用户模型构建模块 300 用于获取文档分类信息，根据文档分类信息构建用户模型。

[0092] 在一个实施例中，根据文档的分类信息构建的用户模型为个人模型和大众模型的线性叠加。其中，个人模型为由用户发表的文档的分类概率组成的向量。如有 n 个分类，那么个人模型就是一个 n 维的向量，其向量的各个元素为用户发表的文档的分类概率。例如，个人模型可以表示为：

[0093] P(people) = (y_1, y_2, …, y_n)

[0094] 其中，y_i 为用户发表的一类文档的分类概率。如设置两个类别，体育类文档和数码科技类文档。用户共发布体育类文档 20 篇，发表数码科技类文档 80 篇。则用户的个人模型为 P(people) = (0.2,0.8)。

[0095] 由于部分用户发表的文档可能很少，其个人模型的可信度并不高，并且在个性化搜索实施初期，很多用户需要一个初步接受的过程，因此需要设置大众模型。大众模型为由

检索词的分类概率组成的向量。由于检索词很短，可采用传统的贝叶斯自动分类方法估算其分类概率。例如，大众模型可以表示为：

[0096] $P(query) = (x_1, x_2, \cdots, x_n)$

[0097] 其中，x_i 为检索词属于文档各分类的概率。如设置了体育和数码科技两个类别，检索词“苹果”是体育类的概率为 0.2，是数码科技类的概率为 0.8，则用户的大众模型为 $P(query) = (0.2, 0.8)$。

[0098] 该实施例中，用户模型构建模块 300 根据个人模型和大众模型构建的用户模型可表示为：

[0099] $P(people\_social) = a \times P(query) + (1-a)P(people)$，$0 \leqslant a \leqslant 1$

[0100] 其中，$P(people\_social)$ 为用户模型，$P(query)$ 为大众模型，$P(people)$ 为个人模型，a 为用户活跃度指数。

[0101] 该实施例中，上述系统还包括用户活跃度指数计算模块（图中未示出）。用户活跃度指数用来判断个人模型的可信度。有的用户仅发表几篇文档，而有的用户发表上千篇文档，用户活跃度指数能够用来综合考虑大众兴趣和个人兴趣。用户发表的文档总数太少，则其兴趣主要以大众兴趣为主，用户发表的文档总数很多，则增加个人模型的权重，因此用户活跃度指数影响个人模型和大众模型的线性叠加参数。

[0102] 在一个优选的实施例中，用户活跃度指数计算模块计算用户活跃度指数的计算公式为：

[0103] $$a = \begin{cases} \dfrac{N}{2 \times N1}, N < 2 \times N1 \\ 1, N \geq 2 \times N1 \end{cases}$$

[0104] 其中，N 为一个用户发表的文档总数，N1 为所有用户平均发表的文档数。即：如果一个用户发表的文档数超过用户平均发表文档数的 2 倍，则用户活跃度指数为 1，认为其个人模型是可信的。如果小于用户平均发表文档数的 2 倍，则计算用户发表的文档数与 2 倍的用户平均发表文档数的比值，作为个人模型和大众模型进行线性叠加计算的参数，计算用户模型。

[0105] 文档模型构建模块 400 用于获取文档分类信息，根据文档分类信息构建文档模型。

[0106] 在一个实施例中，文档模型构建模块 400 构建的文档模型为由文档属于各分类的概率组成的向量。如有 n 个分类，则文档模型就是一个 n 维的向量，其向量的各个元素为文档属于各个分类的概率。例如可以表示为：

[0107] $P(page) = (z_1, z_2, \cdots, z_n)$

[0108] 其中，z_i 为文档属于各分类的概率。对于长文本（如博客等），可采用传统的支持向量机中的逻辑回归方法估算文档的分类概率，对于短文本（如微博等），可采用传统的贝叶斯自动分类方法估算其分类概率。

[0109] 检索模块 500 用于根据用户输入的检索词进行检索，得到检索结果。

[0110] 排序权值计算模块 600 用于根据用户模型和文档模型计算检索结果的排序权值。

[0111] 排序模块 700 用于根据排序权值对检索结果进行排序。

[0112] 在一个实施例中，如图5所示，排序权值计算模块600包括查询单元610、相似度计算单元620和线性叠加单元630，其中：

[0113] 查询单元610用于获取登录用户的用户模型。查询单元610根据登录用户的标识号可查找到登录用户的用户模型。

[0114] 相似度计算单元620用于获取检索结果中每个文档的文档模型及每个文档的作者的用户模型，计算登录用户的用户模型和文档模型的第一相似度，以及计算作者的用户模型与登录用户的用户模型的第二相似度。

[0115] 该实施例中，计算第一相似度即计算用户的用户模型与文档模型的余弦距离，计算公式为：rank1 = sim(P(people), P(page)) = cos<P(people), P(page)>。根据检索结果中每个文档的作者的标识获取到对应的用户模型。计算第二相似度即计算每个文档的作者的用户模型与登录用户的用户模型之间的余弦距离，计算公式为：rank2 = sim(P(user), P(author)) = cos<P(user), P(author)>，其中，P(user)为登录用户的用户模型，P(author)为每个文档的作者的用户模型。

[0116] 在一个优选的实施例中，上述系统还包括专家指数计算模块（图中未示出）专家指数用于第二相似度的计算中，其计算公式为：

[0117] $$\exp(M)=\frac{M}{M1}$$

[0118] 其中，M为用户发表的某类文档的总数，M1为发表该类别文档最多的用户发布的该类文档的数目。

[0119] 在一个实施例中，计算专家指数后，每个文档的作者的用户模型与登录用户的用户模型的第二相似度的计算公式则为：

[0120] rank2 = cos<P(user), P(author)>*exp(M)

[0121] 也可以采用其他方式修正第二相似度的计算，例如，利用专家指数对第二相似度进行加权处理等。

[0122] 线性叠加单元630用于将第一相似度和第二相似度进行线性叠加，得到排序权值。

[0123] 该实施例中，线性叠加单元630按照如下公式计算排序权值：

[0124] rank = b×rank1+(1-b)rank2, $0 \leqslant b \leqslant 1$

[0125] 其中，b为经验参数。

[0126] 在另一个实施例中，用户模型构建模块300包括更新模块310，如图6所示，更新模块310包括设置单元311、统计单元312、折算单元313和用户模型构建单元314，其中：

[0127] 设置单元311用于设置线性衰减函数、统计周期和统计时间段。

[0128] 由于用户模型是按照其发表文档的分类概率构建的，将各时段发表的文档的篇数按照距离当前的时间进行衰减，然后统一计算衰减后的各类别的文档数，根据衰减后的各类别的文档数构建用户模型，则完成了用户模型的更新。

[0129] 统计单元312用于在统计时间段内统计每个统计周期内用户发表的各类文档数。

[0130] 折算单元313用于根据线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数。

[0131] 在一个优选的实施例中，设置单元311设置衰减函数为f(t) = t，表示用户兴趣随

时间线性衰减。设置统计周期为月，12 个月为一个统计时间段。统计单元 312 统计 12 个月每个月发表的各类文档数，折算单元 313 按照距当前的时间进行折算，计算折算后的用户发表的各类文档数。

[0132] 该实施例中，折算单元 313 采用如下公式进行折算：

[0133] $$Z=\sum_{t=1}^{T}(z_t / f(t))$$

[0134] 其中，Z 表示折算后作者发表的某个类别的文档数，t 表示文档发表日期距离当前的时间，zt 表示 t 时段作者发表的某个类别的文档数，f(t) 为衰减函数。

[0135] 例如，一个用户在统计时间段内，第一个月发表体育类文档数 z1，第二个月发表体育类文档数 z2，...，当前的月份，也就是第 12 个月发表体育文档数为 z12，则该用户发布的体育类文档折算后的数目为 $Z=\frac{z1}{12}+\frac{z2}{11}+\frac{z3}{10}+\cdots+\frac{z11}{2}+z12$。

[0136] 用户模型构建单元 314 用于根据折算后的用户发表的各类文档数构建用户模型。

[0137] 以上所述实施例仅表达了本发明的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。因此，本发明专利的保护范围应以所附权利要求为准。

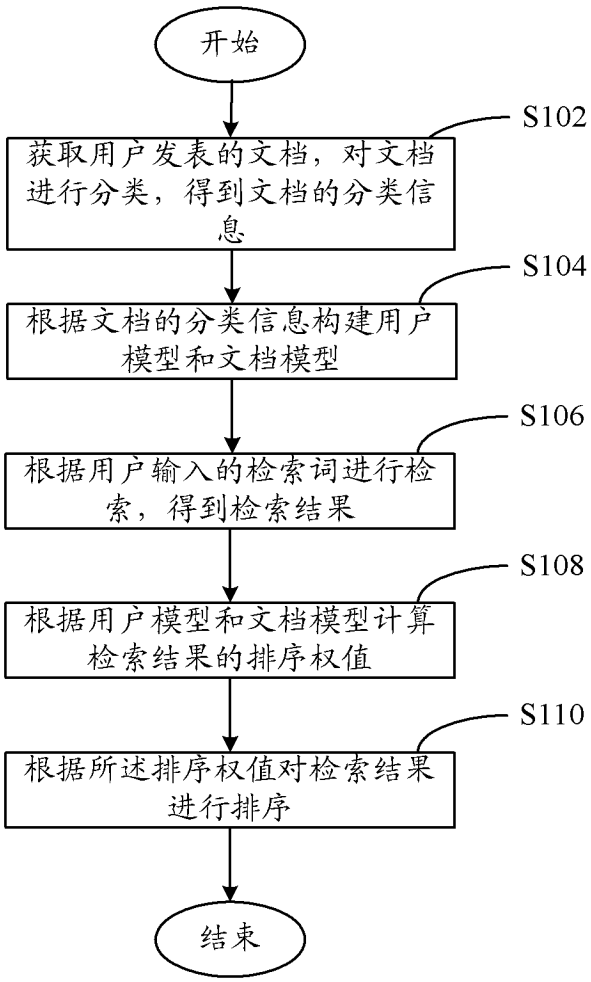

开始
S102
获取用户发表的文档，对文档进行分类，得到文档的分类信息
S104
根据文档的分类信息构建用户模型和文档模型
S106
根据用户输入的检索词进行检索，得到检索结果
S108
根据用户模型和文档模型计算检索结果的排序权值
S110
根据所述排序权值对检索结果进行排序
结束

图 1

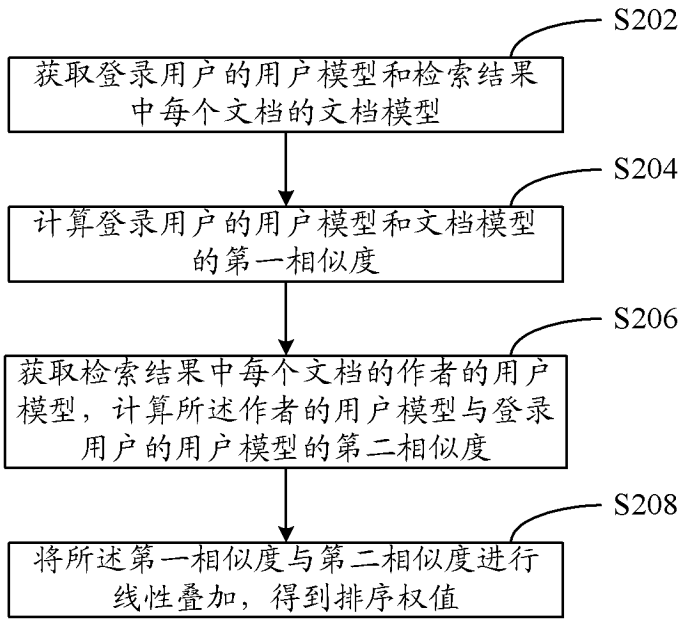

S202
获取登录用户的用户模型和检索结果中每个文档的文档模型
S204
计算登录用户的用户模型和文档模型的第一相似度
S206
获取检索结果中每个文档的作者的用户模型，计算所述作者的用户模型与登录用户的用户模型的第二相似度
S208
将所述第一相似度与第二相似度进行线性叠加，得到排序权值

图 2

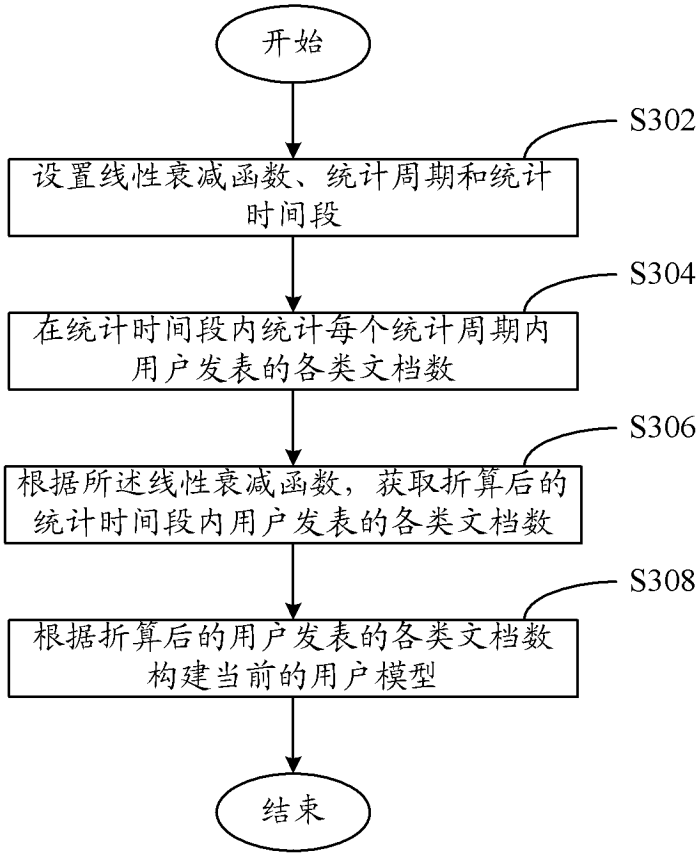

开始
S302
设置线性衰减函数、统计周期和统计时间段
S304
在统计时间段内统计每个统计周期内用户发表的各类文档数
S306
根据所述线性衰减函数，获取折算后的统计时间段内用户发表的各类文档数
S308
根据折算后的用户发表的各类文档数构建当前的用户模型
结束

图 3

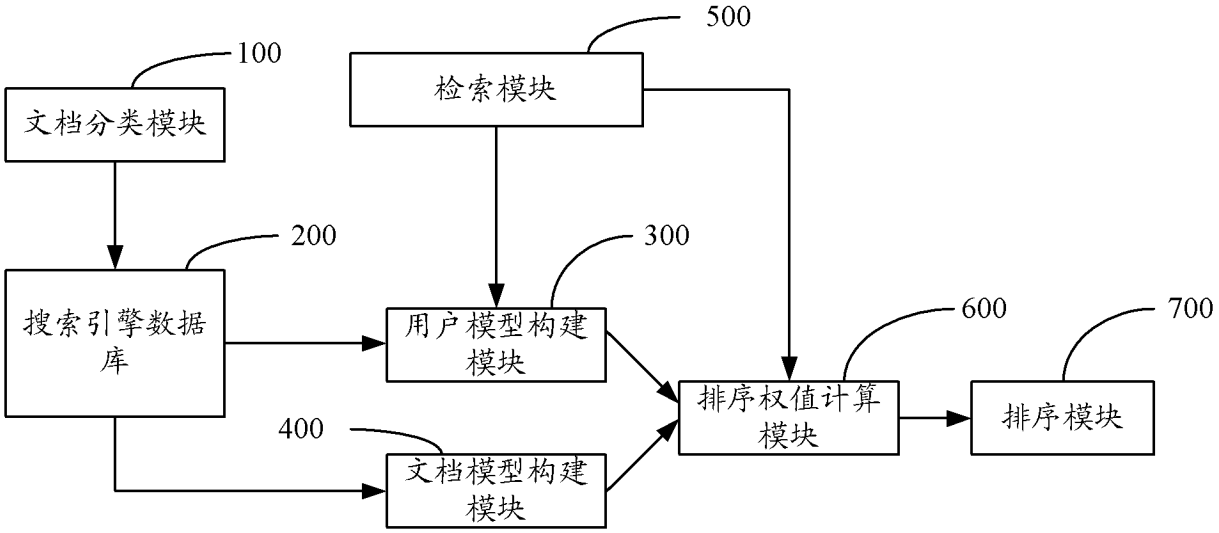

100
文档分类模块
500
检索模块
200
搜索引擎数据库
300
用户模型构建模块
600
排序权值计算模块
700
排序模块
400
文档模型构建模块

图 4

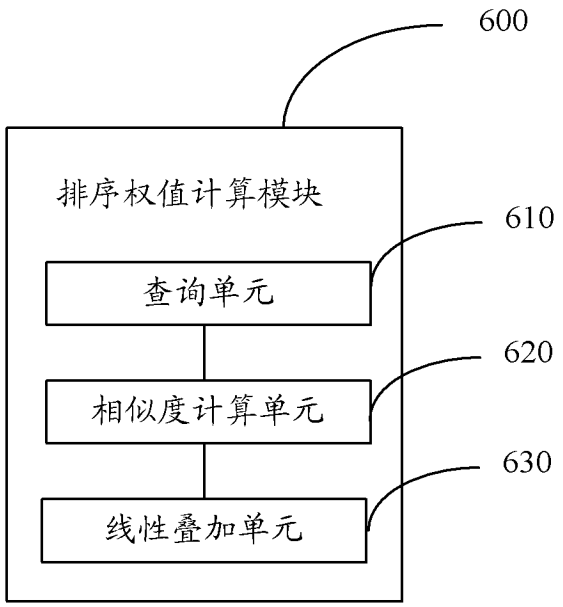

600
排序权值计算模块
610
查询单元
620
相似度计算单元
630
线性叠加单元

图 5

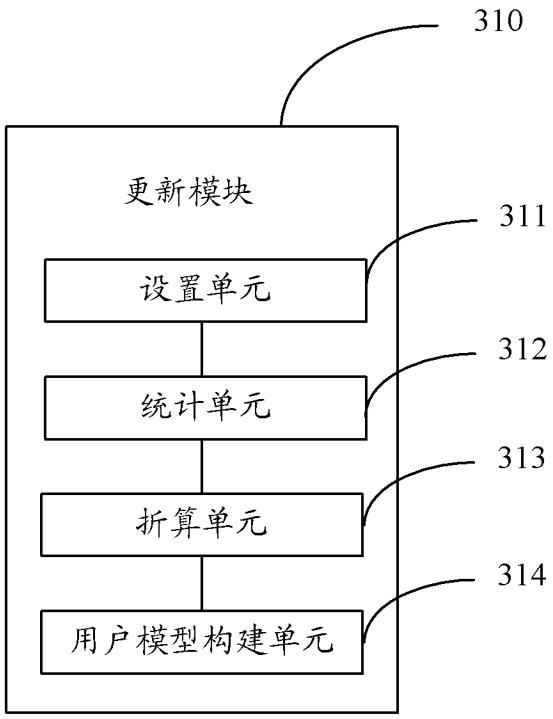

310
更新模块
311
设置单元
312
统计单元
313
折算单元
314
用户模型构建单元

图 6